Bioinformatics Advance Access published April 11, 2012

Fast and accurate inference of local ancestry in Latino populations

Yael Baran [1,*], Bogdan Pasaniuc [2,3,*], Sriram Sankararaman [3,4,*], Dara G. Torgerson [5], Christopher Gignoux [5], Celeste Eng [5], William Rodriguez-Cintron [6], Rocio Chapela [7], Jean G. Ford [8], Pedro C. Avila [9], Jose Rodriguez-Santana [10], Esteban Gonzàlez Burchard [5] and Eran Halperin [1,11,12]

[1]The Blavatnik School of Computer Science, Tel-Aviv University [2]Epidemiology and Biostatistics Dept, Harvard School of Public Health [3]Broad Institute of Harvard and MIT [4]Genetics Dept, Harvard Medical School [5]Bioengineering & Therapeutic Sciences and Medicine Dept, University of California San Francisco [6]Veterans Caribbean Health Care System, San Juan, PR [7] Instituto Nacional de Enfermedades Respiratorias (INER), Mexico City, Mexico [8]Johns Hopkins Bloomberg School of Public Health [9]Division of Allergy-Immunology, Northwestern University [10]Centro de Neumologia Pediatrica, CSP, San Juan, PR [11]Dept. of Molecular Microbiology and Biotechnology, Tel-Aviv University [12]International Computer Science Institute, Berkeley.

Associate Editor: Dr. Jeffrey Barrett

ABSTRACT

Motivation: It is becoming increasingly evident that the analysis of genotype data from recently admixed populations is providing important insights into medical genetics and population history. Such analyses have been used to identify novel disease loci, to understand recombination rate variation and to detect recent selection events. The utility of such studies crucially depends on accurate and unbiased estimation of the ancestry at every genomic locus in recently admixed populations. Although various methods have been proposed and shown to be extremely accurate in two-way admixtures (e.g. African Americans), only a few approaches have been proposed and thoroughly benchmarked on multi-way admixtures (e.g. Latino populations of the Americas).

Results: To address these challenges we introduce here methods for local ancestry inference which leverage the structure of linkage disequilibrium in the ancestral population (LAMP-LD), and incorporate the constraint of Mendelian segregation when inferring local ancestry in nuclear family trios (LAMP-HAP). Our algorithms uniquely combine hidden Markov models of haplotype diversity within a novel window-based framework to achieve superior accuracy as compared to published methods. Further, unlike previous methods, the structure of our HMM does not depend on the number of reference haplotypes but on a fixed constant, and it is thereby capable of utilizing large datasets while remaining highly efficient and robust to over-fitting. Through simulations and analysis of real data from 489 nuclear trio families from the mainland US, Puerto Rico and Mexico, we demonstrate that our methods achieve superior accuracy compared with published methods for local ancestry inference in Latinos.

Availability: http://lamp.icsi.berkeley.edu/lamp/lampld/

Contact: bpasaniu@hsph.harvard.edu

*These authors contributed equally

1 INTRODUCTION

Admixed populations, such as Latinos and African Americans, emerged from the encounter of a few genetically-diverged ancestral populations which have since been mixing for a relatively small number of generations. Due to recombination events, each chromosome of an admixed individual is a mosaic of chromosomal regions originating from the different ancestral populations. The problem of local ancestry inference is to determine, for each genomic position, the ancestral origin of each of the two chromosomes. High-resolution local ancestry inference from genomewide genotype data forms an essential analysis step in medical genetics in identification of disease genes through admixture mapping (Seldin *et al.*, 2011; Hoggart *et al.*, 2004; Reich *et al.*, 2005; Zhu *et al.*, 2004) as well as in increasing power in association studies in admixed populations (Pasaniuc *et al.*, 2011). Local ancestry inference is also useful in the study of population genetic processes, such as recombination (Hinch *et al.*, 2011; Wegmann *et al.*, 2011), selection (Tang *et al.*, 2007) and migration (Bryc *et al.*, 2010), thus providing important insights into human history and demographics. In addition, ancestry inference has been recently shown to be of critical value in pharmacogenomics: A recent study associated the Native American ancestry with the risk of relapse in children suffering from acute lymphoblastic leukemia (Yang *et al.*, 2011).

A number of methods for inferring local ancestry have been proposed (Price *et al.*, 2009; Sankararaman *et al.*, 2008; Tang *et al.*, 2006; Sundquist *et al.*, 2008; Pasaniuc *et al.*, 2009b,a) and have been shown to be very accurate in African Americans (Seldin *et al.*, 2011). Unfortunately, the accuracy of these methods is limited for more complex populations such as Latinos, which are formed by the admixture of three ancestral populations (European, African and Native American). Moreover, the Native American and European ancestries are genetically closer than Africans

and Europeans, making the inference of local ancestry a more challenging task in Latinos and in other populations of similar admixture characteristics (Pasaniuc *et al.*, 2009b). With some of the methods being completely unable to directly handle multi-way mixtures and the rest prone to these difficulties, the error rates of local ancestry estimates in such populations are high and the results of current studies involving Latino populations are hard to interpret. It is therefore crucial to develop highly accurate and efficient methods for local ancestry inference in multi-way admixed populations in conjunction with a comprehensive assessment of their performance.

The critical importance of this problem is further underlined by a number of recently proposed approaches for local ancestry inference in multi-way admixtures developed in parallel to our work (Henn *et al.*, 2012; Johnson *et al.*, 2011; Bercovici *et al.*, 2012). Johnson *et al.* (2011) use an extension of Saber to three-way mixtures in a haploid mode to infer virtual genomes, while Henn *et al.* (2012) extended on the work of Bryc *et al.* (2010) to employ PCA with a post-processing HMM to call local ancestry in multi-way mixtured populations. Bercovici *et al.* (2012) extend on previous work in the context of variable length HMMs for local ancestry inference.

To the best of our knowledge, our approach is the only one capturing the ancestral haplotype structure using a fast approximation of the Li and Stephens (Li and Stephens, 2003) model within a generative framework for admixed genomes, resulting in an efficient and robust algorithm. We note that an exhaustive comparison with the methods that were independently developed recently is beyond the scope of this paper.

We introduce here two methods, LAMP-LD and LAMP-HAP, for local ancestry inference in multi-way admixed populations. Similar to methods proposed for African Americans (Price *et al.*, 2009; Sundquist *et al.*, 2008), our methods leverage the haplotype structure of the ancestral populations to infer local ancestry. While HAPMIX and HAPAA (Price *et al.*, 2009; Sundquist *et al.*, 2008) model this haplotype structure using Hidden Markov Models (HMMs) (Li and Stephens, 2003) with state space and runtime quadratic in the number of reference haplotypes, the haplotype model underlying our methods employs HMMs with a fixed-size state space (Scheet and Stephens, 2006; Kennedy *et al.*, 2008; Kimmel and Shamir, 2005). As a first advantage of this approach, the running time of our algorithms is an order of magnitude faster than previous HMM-based methods such as HAPMIX (Price *et al.*, 2009). In addition, our model estimates its parameters from the reference haplotype data, and therefore is less prone to inaccuracy due to misspecification of parameters than models that use the standard Li and Stephens HMM approach (Li and Stephens, 2003), in which certain parameters such as the population recombination rates are required as input.

Another important feature of our algorithms is the integration of the HMM within a window-based framework. It has been noted that a straightforward extension of the standard Li and Stephens model to admixed chromosomes would tend to predict artificially frequent transitions in local ancestry (Price *et al.*, 2009). This effect arises due to the limited sample size of the reference panels, as some ancestral haplotype segments in the admixed population may not be represented in the reference panels. Methods such as HAPMIX (Price *et al.*, 2009) mitigate this effect by introducing a ”miscopying” parameter that summarizes the probability of miscopying of haplotype segments among ancestries. In our approach, we solve this problem by dividing the genome into non-overlapping windows such that no transitions between ancestries are made within each window, an assumption we relax in a post-processing stage. Limiting the occurrence of ancestry transitions in this window-based framework, together with the ”fuzziness” of our HMMs, greatly improves the inference quality by eliminating extremely short, likely artifactual, ancestral segments.

The approach described above is implemented by the method LAMP-LD for inferring local ancestry in both genotype and haplotype data of unrelated individuals. However, in studies of admixed populations it is often the case that multiple family members are genotyped. The availability of such pedigree information could be leveraged to further improve estimates of local ancestry. Hence, we developed LAMP-HAP, an extension of LAMP-LD, to infer local ancestry in nuclear family trios.

Using extensive simulations, we show that LAMP-LD provides a substantial improvement in the accuracy as well as efficiency of local ancestry inference in Latinos over published approaches. We also show that LAMP-HAP achieves increased accuracy over LAMP-LD, thus demonstrating the utility of integrating family information in local ancestry inference.

In practical applications, a number of key questions need to be addressed to enable accurate local ancestry inference in Latinos. Unlike African Americans that are well modeled by a mixture of West Africans and Europeans, it is currently unclear which combination of reference haplotypes optimizes local ancestry inference in Latinos (Seldin *et al.*, 2011); it is therefore critical to assess the effect of a proxy reference haplotype set on accuracy. First, in light of LAMP-LD’s ability to efficiently handle large reference sets, we assess whether increasing the size of the reference data results in superior accuracy. Second, we examine LAMP-LD’s sensitivity to the genetic divergence between the ancestral haplotypes and the proxy data used as reference. We observe that LAMP-LD successfully translates the increased size of the reference set as well as the lower divergence between proxy and ancestral population into superior accuracy.

We also present an evaluation of the effect of European gene flow into present day Native American populations on local ancestry inference in Latinos. This effect is important to assess as it is estimated that most Native American populations used as reference panels in local ancestry inference have been exposed to European gene flow. Through simulations we find that the presence of European segments in the Native American reference panels yields biased local ancestry estimates. However, our results suggest that under a small amount of gene-flow (under 6%) these effects would yield statistically significant association in case-only admixture studies only at very large sample sizes.

We conclude by assessing the performance of our methods on real Latino data. Testing local ancestry inference in empirical data is important, since simulations of admixed genotypes inherently make assumptions about the mixture process, (e.g. number of generations, per-generation mixture proportions, the availability of ancestral haplotypes) which do not necessarily hold in the analysis of real data. Here, we use 489 Mexican and Puerto Rican trio families from the Genetics of Asthma in Latino Americans (GALA) (Burchard *et al.*, 2004) to estimate local ancestry performance. We use the fact that the true local ancestry along the chromosome follows Mendelian inheritance rules, and thus count the Mendelian inconsistencies in the local ancestry (MILANC) estimates produced

by methods that treat every sample in the family as unrelated. We find that our method attains lower MILANC rates, thus establishing its superior accuracy in an empirical setting.

2 METHODS

We model recently admixed chromosomes as a set of haplotypes from K ancestral populations that have come together at some point in time and have been mixing through random mating for g generations. Formally, we note by $\alpha = (\alpha_1, \alpha_2, \cdots \alpha_K)$ the fraction of haplotypes from each of the K ancestral populations at the time of the encounter. After g generations, each chromosome can be modeled as a random walk from the $5'$-end to the $3'$-end, with crossovers between chromosomes occurring as a Poisson process with rate $g\rho$, where ρ is the average recombination rate across the genome. The recombination events break the ancestry and insert ancestry switches (also called *breakpoints*). Conditional on the positions of such switches, each segment between two consecutive breakpoints is modeled as an independent draw from the ancestral populations with probabilities given by the admixture fraction α. For simplicity of exposition we describe our methods and simulations assuming a constant recombination rate, however we note that they can be easily adjusted to account for position-specific rates by scaling the physical positions of the SNPs with any specific recombination map.

Briefly, our model consists of a top level HMM which emits genotypes in non-overlapping windows. The hidden states of this HMM correspond to the local ancestries on each chromosome within each window (we initially restrict changes in local ancestry to the window boundaries). Given the pair of ancestral states within a window, the genotypes are emitted by a pair of sub-HMMs which model the corresponding ancestral populations. The parameters of the sub-HMMs are estimated from the reference panels for these populations. Given the parameters, we compute the most likely pair of local ancestries in each window, followed by a post-processing step which relaxes the restriction on the localization of ancestry switches.

2.1 Modeling linkage disequilibrium (LD) in the ancestral populations

Several approaches have been proposed for local-ancestry estimation in two-way admixtures, with methods which explicitly model the linkage disequilibrium (LD) structure within the ancestral populations showing the highest accuracy in African Americans (Pasaniuc *et al.*, 2009a; Price *et al.*, 2009; Sundquist *et al.*, 2008). These methods can be broadly classified under two main approaches according to the type of the HMM which models the LD. The methods proposed in Price *et al.* (2009) and in Sundquist *et al.* (2008) use HMMs with state space and runtime quadratic in the number of reference haplotypes. Therefore, these methods are impractical for large sets of reference haplotypes, as is the case for multi-way mixtures; for example, HAPMIX (Price *et al.*, 2009) takes 7 hours to perform local ancestry inference in a sample of 200 African Americans on chromosome 1 when HapMap European and African haplotypes are used as reference (Seldin *et al.*, 2011). This raises the need for scalable and accurate methods for local ancestry inference that are capable of handling the ever-growing number of reference haplotypes. The second class of HMMs aims to achieve this through a fixed state space (described by a constant S) independent of the size of the reference panels. So far, only one method has attempted at using such HMMs in the context of local ancestry inference, namely GEDI-ADMX (Pasaniuc *et al.*, 2009a). Unlike GEDI-ADMX that uses an ad-hoc metric (imputation accuracy) requiring masking and re-imputation of every SNP genotype in the data to infer ancestry, leading to increased runtime, here we extend fixed-structure HMMs into a fully generative model for admixed chromosomes within a non-overlapping window-based framework. This leads to superior accuracy in simulations of Latinos (see Results section).

The structure of our model is fully described by a constant S and a window length L. There are $S \times L$ states in our model, with each state emitting the reference or alternate allele according to an emission probability ϵ. Any haplotype (over the L SNPs) can be generated across any path of L states according to the transition and emission probabilities in the model. These probabilities are directly estimated from the reference haplotype data using the Baum-Welch algorithm. We learn HMMs for each of the ancestral populations, and these HMMs are then used for local ancestry inference, as described in Section 2.2. Intuitively, our model "compresses" the diversity observed across all the reference panel within a set of S prototypical states at each SNP (typically much smaller than the number of reference haplotypes).

Formally, the HMM is specified by a triple $M = (Q, \delta, \epsilon)$, where Q is the set of states, δ is the transition probability function, and ϵ is the emission probability function. The set of states Q consists of disjoint sets $Q_0 = \{s_0\}, Q_1, Q_2, \cdots, Q_L$, with $|Q_1| = |Q_2| = \cdots = |Q_L| = S$, where L denotes the set of SNPs, s_0 denotes the start state and Q_i denotes the set of states corresponding to SNP i. $\delta_j(s, s')$ denotes the transition probability of moving from state s at SNP j to state s' at SNP $j+1$, such that $\sum_{s'} \delta_j(s, s') = 1$. The initial state is silent while each other state s emits the reference with probability $\epsilon_j(s, 1)$ and the alternate allele with probability $\epsilon_j(s, 0) = 1 - \epsilon_j(s, 1)$. The probability of observing a haplotype $H = H_1 H_2 \cdots H_n$ given the model M is given by:

$$P(H|M) = \sum_{\pi} \delta_0(s_0, \pi_1)\epsilon_1(\pi_1, H_1) \prod_{i=2}^{L} \delta_i(\pi_{i-1}, \pi_i)\epsilon_i(\pi_i, H_i) \quad (1)$$

where the sum is taken across all paths of states $\pi = \pi_1 \cdots \pi_n$. The summation can be efficiently computed in time $O(S^2 L)$ using the standard HMM forward-backward computations.

Intuitively, a larger S induces a better modeling of the haplotype structure with significant increase in run time. By fixing S to a moderately small number, we achieve large improvements in run time with very modest reductions in accuracy. In contrast to the standard model of Li-Stephens, we estimate the transition and emission probabilities directly from the haplotype data available for each ancestral population. When high quality maps of recombination rates are available, it would be beneficial to use the known recombination rates instead of learning those from the data, however it is often the case that the recombination maps have poor quality, particularly if the proxy populations do not accurately represent the true ancestral populations. In addition, the parameters of our model can be estimated using genotype data directly (Kennedy *et al.*, 2008; Kimmel and Shamir, 2005), thus making the model robust to phasing errors.

2.2 A window-based framework for local ancestry inference

We use the above HMM as a building block for a window-based HMM, as we now describe. We divide the genome into non-overlapping windows $w = [i, i + L)$ of length L, spanning SNPs i to $i + L$. Within each window we make the assumption that no breakpoints (crossovers that change ancestry) occur and thus we constrain all breakpoints to occur at the boundary of two consecutive windows. We will show below how this assumption can be relaxed in a post-processing step of the algorithm. We train the HMM of the ancestral population separately for each window (particularly, there is a separate start state for each window). Therefore, our model for representing admixed chromosomes can be viewed as a top level HMM with $\binom{K}{2}$ states corresponding to each pair of ancestries $S^w = \{(M_1^w, M_2^w)\}$, and for each window $w = [i, i + L)$ across the genome(see Figure 1). Each state (M_1^w, M_2^w) emits a genotype block G^w with emission probability:

$$\sum_{(H_1^w, H_2^w)} P(H_1^w|M_1^w)P(H_2^w|M_2^w)$$

where $P(H_1^w|M_1^w)$ is the probability of emitting the haplotype segment H_1^w under the HMM for ancestry M_1 (Equation 1) and (H_1^w, H_2^w) is a pair of haplotypes that is compatible with G^w. This probability can be efficiently computed using standard extensions of forward-backward algorithms in time $O(S^4L)$. We implemented the factorization speed-up of (Kennedy *et al.*, 2008) to achieve a running time of $O(S^3L)$ for computing the probability of a genotype over L SNPs given the model M.

The transition probability from state (M_1^w, M_2^w) to state $(M_1^{w'}, M_2^{w'})$, where $w' = [i+L, i+2*L)$, is set to $\theta = 10^{-8} \times D$ (D is the length in base-pairs between windows) if unordered ancestry pairs (M_1^w, M_2^w) and $(M_1^{w'}, M_2^{w'})$ differ by one ancestry, θ^2 if both ancestries differ and 1-2θ -3θ^2 if the respective ancestry pairs are the same.

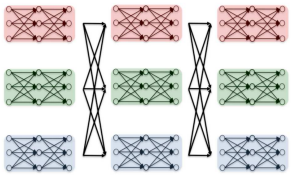


Fig. 1. Schematic structure of our model (haploid version is displayed for simplicity) over 9 SNPs with 3 windows each of length 3 SNPs. In each window, the haplotypes of each ancestral population are modeled using distinct HMMs (denoted in different colors). Transitions that change the ancestral population are allowed only at the boundary of consecutive windows. This framework is generic in that any model (e.g. Li and Stephens(Li and Stephens, 2003), fastPHASE (Scheet and Stephens, 2006)) can be used to account for the ancestral LD.

Decoding within the top level HMM is performed using a standard Viterbi decoding, a dynamic programming algorithm that runs in time proportional to the number of windows and squared in the number of states $\binom{K}{2}$. The Viterbi decoding gives an ancestry assignment to each window constraining all the breakpoints to occur at the boundaries of each window. To account for this rather strict assumption, in the final step of the algorithm, we consider all the breakpoints identified in the second stage and find a high-resolution localization for each of them as follows. For a breakpoint that occurred between window i and $i + 1$, we use a simplified version of the window-based HMM (allowing only one breakpoint) to infer the ancestry of the genotype in windows i and $i + 1$. Using the HMMs of the ancestral populations we compute the probability of the observed haplotype given the breakpoint occurring in any of the SNP positions in the three windows and pick the location that maximizes this likelihood. This computation is achieved in time proportional to the size of the three windows spanning the inferred breakpoint.

2.3 Incorporating trio information in local ancestry inference

We denote a nuclear family (trio) as a triplet of vectors (g^M, g^F, g^C), each of size n, corresponding to genotypes over n typed SNPs. The genotypes $g_j^i \in \{0, 1, 2\}$ are the counts of the reference allele at SNP j in individual i. Due to Mendelian inheritance rules, in every trio at every SNP j there are four independent alleles: the maternal transmitted and un-transmitted alleles, and the paternal transmitted and un-transmitted alleles. We are interested in estimating the ancestral population of each of the four alleles at each SNP in each trio.

The full HMM described above can be extended to trios using standard factorial HMM by jointly modeling the parental transmitted and un-transmitted haplotypes conditional on the observed trio genotype data. However, this approach is impractical for large scale datasets due to the joint modeling of four paths in the HMM (Kennedy *et al.*, 2008). Here, we take a two step approach to performing local ancestry inference in trios. First we use the trio genotypes to perform phasing and obtain estimates of transmitted and un-transmitted haplotypes (data at trio-ambiguous SNPs is set to missing). Second, we estimate the local ancestry in each of the four haplotypes independently using a haploid version of the model described above. We show in the Results section that this two step approach produces accurate results, comparable to what would be achieved when the true phasing is known, thus showing that an approximate phasing of the data using the trios is as useful as perfect phasing for the purpose of local ancestry inference. More importantly, we show that the accuracy of local ancestry inference is considerably improved in trios compared to unrelated individuals.

3 RESULTS

Latino populations of the Americas, such as Mexicans or Puerto Ricans, arose by the influx of Europeans into existing Native American populations. Subsequently, African individuals were introduced into the population (Tang *et al.*, 2007; Morales Carrión, 1983). Thus, most of the genomes of current Latino populations can be modeled as an admixture of chromosomes from three ancestral populations with various global proportions of European, Native American and West African ancestries (*e.g.*, .45:.5:.05 for Mexicans and .67:.13:.2 for Puerto Ricans (Price *et al.*, 2007; Burchard *et al.*, 2005; Mao *et al.*, 2007; Tian *et al.*, 2007)). Correspondingly, we simulated Latino admixed haplotypes as mosaics of segments taken from three of the HapMap phase 3 haplotype panels (The International HapMap Consortium, 2005). Unless otherwise noted,

we used the phased haplotypes from the CEU (117 haplotypes), CHB+CHD (169) and YRI (115) panels in our simulations of admixed haplotypes and phased haplotypes from the TSI (117), JPT (169) and LWK (115) panels as proxy reference data in local ancestry inference. The haplotype sets used for generating the simulations data and reference data are therefore disjoint. Our use of East Asian haplotypes to represent the Native American haplotypes was motivated by the small sample sizes of existing Native American panels and by the presence of European gene flow into some of these populations. It is likely, however, that the use of East Asian haplotypes will overestimate the accuracy of local ancestry inference.

We performed the analyses on chromosome 10, restricted to the SNPs present on the Illumina Human 1M SNP array so as to obtain a realistic SNP density and a typical genomic LD pattern. Following standard approaches (Price *et al.*, 2009), we simulated admixed chromosomes by performing a random walk over the HapMap haplotypes. Distance to the next crossover was sampled from the exponential distribution with parameter $\frac{1}{\theta g}$, where $\theta = 10^{-8}$ is the average recombination probability along the genome per base per generation, and $g = 15$ is the approximate number of generations in admixture for Latinos. At a crossover event the new ancestry is chosen given the mixture-specific proportions, and a specific haplotype is drawn uniformly from the corresponding reference set. This procedure was used to generate 400 haplotypes, which were then joined in pairs to form 200 diplotypes.

Several metrics have been proposed to measure the performance of local ancestry inference methods (Seldin *et al.*, 2011). Here we use the squared Pearson correlation coefficient r^2 between the true and the inferred number of alleles from each of the ancestries, averaged over the three ancestries. The squared correlation is directly related to the power achieved in case-only admixture mapping, *i.e.*, N/r^2 cases are required to achieve the same power as a study with N cases where the local ancestries are known without error (see Supplementary Note). The second measure we use is the percentage of all SNP loci whose diploid ancestry was incorrectly inferred, which we refer to as the *Diploid Error*.

3.1 Comparison with other methods

Several methods have been developed for inferring local ancestry (Sundquist *et al.*, 2008; Tang *et al.*, 2006; Price *et al.*, 2009; Sankararaman *et al.*, 2008; Pasaniuc *et al.*, 2009b,a; Patterson *et al.*, 2004; Bryc *et al.*, 2010) and have been shown to attain very high accuracy in admixtures of two genetically diverged ancestral populations such as African Americans (Price *et al.*, 2009; Pasaniuc *et al.*, 2009b). Only a few of these methods have been extended to admixtures of three populations such as Latinos (Pasaniuc *et al.*, 2009a,b; Henn *et al.*, 2012; Johnson *et al.*, 2011; Bercovici *et al.*, 2012), and we compared LAMP-LD to two of them. The first is WINPOP, a method shown to attain high accuracy in simulated data (Pasaniuc *et al.*, 2009b), which has been used in a number of recent empirical studies of Latinos (Bryc *et al.*, 2010; Yang *et al.*, 2011). WINPOP treats the observed genotypes as independent given the local ancestry, thereby ignoring the haplotype structure within each population. The second is GEDI-ADMX, which is similar to our approach in using fixed size HMMs to model haplotype diversity, but uses a completely different framework for inferring ancestries at each locus in the genome. We also compared LAMP-LD to HAPMIX (Price *et al.*, 2009). Although LAMP-LD and HAPMIX are similar in that they require reference haplotypes from each of the ancestral populations, the HMMs employed by the two models have different structure. In addition, LAMP-LD traverses the chromosome using the window-based framework, while HAPMIX employs a "miscopying" parameter to account for imperfections in the reference panels.

As a safety check, we first simulated 2-way mixtures of African Americans using .8:.2 proportions for YRI and CEU respectively with 6 generations of admixture. On this data LAMP-LD attained an average $r^2 = 0.99$, very similar (no significant difference) with the $r^2 = 0.98$ attained by HAPMIX, thus confirming the high accuracy of local ancestry inference in African Americans (Price *et al.*, 2009; Seldin *et al.*, 2011).

Since HAPMIX was not designed to directly process multi-way mixtures, we adapted it to the task by running it two times on each genotype. The first run aimed at discerning the African segments from the rest of the segments: One reference panel included the TSI and the JPT haplotypes, and the other one comprised the LWK haplotypes, with the mixture proportion set to the proportion of the African ancestry in the mixture. The second run aimed at discerning between the European and the Native American segments. For the Mexican simulations, the first reference panel included the TSI haplotypes, the second panel included the JPT haplotypes, and the mixture proportion was set to the relative share of the European and Native American ancestries in the non-African segments. For the Puerto Rican simulations, the first reference panel included TSI+LWK haplotypes, the second panel included the JPT haplotypes, and the mixture proportion was set to the proportion of the Native American ancestry in the mixture. The different schemes were designed to account for the fact that the proportion of African ancestry is small in Mexican data (5%) but considerably higher in the Puerto Rican data (20%), and were matched to the datasets as to yield more accurate results. Throughout the paper we denote the described schemes for running HAPMIX jointly as HAPMIX*.

Method	Mexican		Puerto Rican	
	% diploid error	r^2	% diploid error	r^2
WINPOP	12.8 (0.3)	0.804	9.0 (0.3)	0.817
GEDI-ADMX	16.9 (0.3)	0.693	13.3 (0.3)	0.723
HAPMIX*	12.9 (0.4)	0.802	16.3 (0.4)	0.697
LAMP-LD	**9.9** (0.3)	**0.847**	**6.4** (0.2)	**0.868**

Table 1. Accuracy (standard error of the mean) attained by the compared methods averaged over 200 simulated Latino genotypes. Diploid error is averaged over genotypes, r^2 is averaged over the three ancestries. HAPMIX* denotes our adaptation of HAPMIX to three-way mixtures. LAMP-LD uses $L = 50$ and $S = 10$ as default parameters (see Section 3.2). LAMP-LD yields the highest accuracy as measured by both metrics on both Mexican and Puerto Rican simulations.

Table 1 compares LAMP-LD to WINPOP, GEDI-ADMX and HAPMIX* on the Mexican and on the Puerto Rican datasets. LAMP-LD achieves the highest accuracy under both the r^2 and the diploid error on both datasets, showing a considerable improvement compared with WINPOP, thus reflecting the utility of the LD information. HAPMIX* attains comparable accuracy with

WINPOP in the Mexican simulations and much worse on the Puerto Rican data; this could be because the parameters of the HAPMIX model were not optimized for Latinos - for example, it is not obvious how to set the effective population size parameter for HAPMIX in these scenarios. However, we should note that HAPMIX was not designed for multi-way mixtures and it could potentially be improved by a more principled extension to multi-way mixtures.

In addition to its high accuracy, LAMP-LD runs an order of magnitude faster compared with HAPMIX. Each run of LAMP-LD is composed of a preliminary stage in which the HMMs are constructed from the reference panels and a second stage of actual inference on the given genotypes. In the experiments above the first stage took 56 minutes, and the processing of each genotype 6.5 seconds (all running times were measured on a single AMD Opteron 1.1 GHz processor). These numbers can be used to extrapolate the running time over 1,000 genotypes, obtaining $\sim$ 3 hours for chromosome 10. HAPMIX's runtime, on the other hand, is linear in the number of genotypes, requiring 89 seconds for each. Running it on 1,000 genotypes would therefore require over 24 hours for one chromosome. This leads to a runtime of ~ 3 days for a full genome scan for LAMP-LD as compared to over $\sim$ 22 days for HAPMIX on a single CPU.

3.2 Assessment of model parameters

The only two parameters required by LAMP-LD are the number of states S and the window length L . We assessed the performance of our method as a function of these parameters. Figure 2 shows that the accuracy is maximized at a value of $L = 50 - 100$ SNPs, corresponding to roughly 200 to 400 Kb on average in our simulated chromosomes. Interestingly, the optimal value for L is fairly stable across the two different populations, suggesting that this parameter can be set independently of the specific mixture proportions. We note that although these results are likely to be specific to the SNP density in our datasets (a SNP every 4,350 bases on average), increasing L above 500 to accommodate for denser SNP panels has only a minor effect on the running time.

Next, we assessed the robustness of our method to different values of S, the number of states per SNP (L is set to 50). The results are presented in Figure 2. As expected, the diploid error decreases as S increases; however, increasing S above 10 provides only marginal improvement in accuracy, reflecting the fact that most of the haplotypic diversity within the reference panels necessary for accurate local ancestry inference is captured by 10 states per population. This is especially important because the running time of HMM-based methods increases quadratically with the number of states. This advantage of LAMP-LD is reflected in the large differences in running time between LAMP-LD and HAPMIX presented in section 3.1, since in order to utilize the entire reference set HAPMIX employed nearly 400 states, each modeling a single reference haplotype. According to the results of this section, if not explicitly noted, all results of this paper for LAMP-LD use parameters $L = 50$ and $S = 10$.

3.3 Advantage of incorporating trio information in local ancestry inference

We simulated nuclear family trios by generating one offspring haplotype from each of the 200 simulated admixed genotypes, followed by grouping the offspring haplotypes into 100 pairs, each forming the genotype of a single progeny. An offspring haplotype was generated by recombining the two parental haplotypes according to the average genomic recombination rate. We then compared the performance of LAMP-LD and LAMP-HAP when inferring local ancestry in the Mexican and Puerto Rican datasets assuming different amounts of information in the inference. For consistency the accuracy was assessed only on the parental genotypes for both methods. Additionally, we measured the accuracy of LAMP-HAP when the haplotype phase is known (*i.e.*, the method receives as an input the true phasing for the simulated trio data) so as to provide an upper bound on the achievable accuracy using trio data.

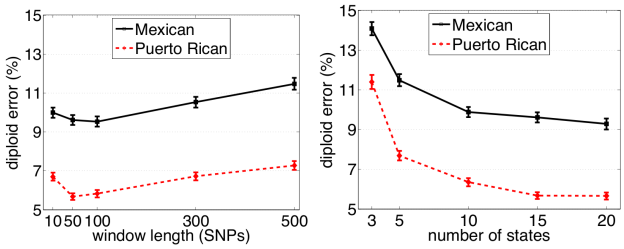


Fig. 2. Effect of the window length (left) and the number of states parameter (right) on accuracy of LAMP-LD. We observe that a window length of 50 -100 SNPs (200-400 Kb) minimizes the error rate for both simulations. Accuracy increases with the number of states S, however, 10-15 states are sufficient for capturing most of the ancestral genetic variation for the purpose of local ancestry inference.

Method	Mexican		Puerto Rican	
	% diploid error	r^2	% diploid error	r^2
LAMP-LD	9.9 (0.3)	0.847	6.4 (0.2)	0.868
LAMP-HAP	**6.6** (0.2)	**0.885**	**5.3** (0.2)	**0.891**
LAMP-HAP*	6.1 (0.2)	0.892	5.0 (0.2)	0.897

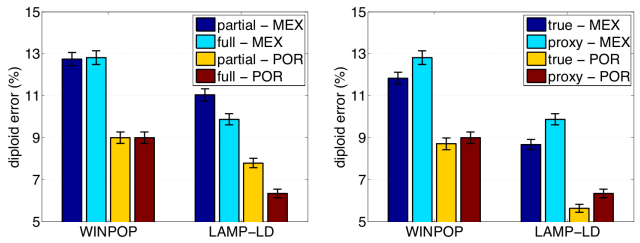
Table 2. Error rate (standard error of the mean) of methods for local ancestry inference as a function of the amount of information taken into account. LAMP-HAP* is provided the true haplotypes used to simulate the trio, as to provide an upper bound on the accuracy that can be achieved in trio data.

The result in Table 2 show a considerable increase in the accuracy, as measured by the diploid error as well as by the squared correlation, with the incorporation of pedigree information. Interestingly, only a marginal improvement was obtained when we provided the true haplotypes to LAMP-HAP, demonstrating that the unambiguously phased positions are sufficient for highly accurate ancestry inference.

3.4 The effect of size and precision of reference sets on accuracy

Most local ancestry inference methods require some information about the mixing populations: Haplotype-based methods, such as LAMP-LD and HAPMIX, require sample haplotypes, while other methods, such as WINPOP, require only SNP allele frequencies. With the growing availability of genetic data, it is important

to examine the effect of the reference datasets (genotypes or haplotypes) on the performance of the methods. Particularly, since LAMP-LD is able to efficiently process large reference datasets, an interesting question is whether it can utilize the additional information provided in sets of growing sizes, given the fact that it uses only 10 prototype haplotypes (states) per ancestry.



(a) Reference Size (b) Reference Similarity

Fig. 3. Effect of reference panel size and divergence on the accuracy of WINPOP and LAMP-LD. Both methods show increased performance with sample size with LAMP-LD showing the highest gain in accuracy when more accurate reference haplotypes are provided as proxy panels.

This question was tested by providing LAMP-LD with reference sets of varying sizes: We compared the results obtained on the *full* set used in the previous sections (117 TSI haplotypes, 169 JPTs and 115 LWKs) to those obtained on a *partial* reference, which contained only two thirds of the haplotypes in each of the three ancestral panels. We did the same with WINPOP, to examine how a non haplotypes-based method would be affected. In Figure 3(a) LAMP-LD can be seen to considerably improve when provided with the larger reference. In contrast, WINPOP does not improve, presumably because estimating the allele frequencies can be done well enough using small panels. On the other hand, LAMP-LD's performance also deteriorates more rapidly as the reference size decreases, and WINPOP's accuracy becomes superior when using 0.4 and 0.5 of each panel for the Mexican and Puerto Rican datasets, respectively (these fractions correspond to panel sizes of 46/58 TSI haplotypes, 67/84 JPTs and 46/57 LWKs).

It has been shown that the genetic divergence between the haplotypes used as proxy and the true unknown ancestral population greatly impacts local ancestry performance (Pasaniuc *et al.*, 2009b; Price *et al.*, 2009). We quantified this effect in Latinos by running LAMP-LD and WINPOP using the *proxy* reference set (which included haplotypes from the TSI, JPT and LWK panels; the same populations were used to obtain the previous results presented in this paper) and a *true* reference set. The true reference in this experiment included the same number of haplotypes in each ancestral panel as the proxy set, but taken from the CEU, (CHB+CHD) and YRI panels; we note that the haplotypes in this set are different from those used for generating the simulated haplotypes.

Figure 3(b) demonstrates the anticipated deterioration in the performance of both methods on both datasets when data from the proxy populations is used as reference instead of the true ancestral populations. This decrease is smaller on the Puerto Rican dataset, presumably because it contains a larger proportion of African ancestry which is more easily differentiated from the rest, even when the proxy LWK haplotypes are used. The deterioration in accuracy is at the same scale as the improvement resulting from increasing the reference size, suggesting that a large enough reference would compensate for the divergence.

3.5 The effect of European gene flow into Native American reference haplotypes

Current day Native American haplotypes used as proxy for the Native American component of Latinos are presumed to contain European gene flow. In order to test the effect of this phenomenon on ancestry inference, we introduced TSI segments into the Asian haplotypes of a reference set composed of 117 CEU, 169 (CHB+CHD) and 115 YRI haplotypes. We performed 10 experiments, in each choosing at random a 5 Mb region along the chromosome, and replacing a percentage of the (CHB+CHD) haplotypes with TSI haplotypes along the chosen region.

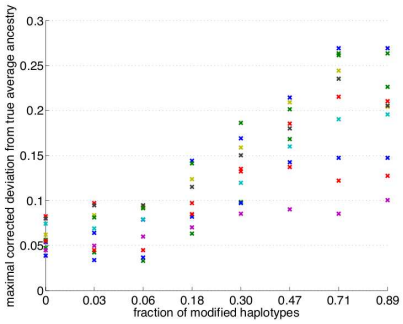


Fig. 4. The simulated effect of European gene flow into Native American haplotypes on estimation accuracy. Different fractions of the 169 (CHB+CHD) reference haplotypes were replaced by TSI haplotypes along 10 different regions, each of length 5 Mb. The plot gives, for each fraction and for each region (in different colors), the maximal corrected deviation (see text for details) of the estimated average European ancestry from the true average european ancestry along the region.

We observed that the typical effect of increasing the number of TSI segments present in the Native American reference panels is an increase in the estimated proportion of the Native American ancestry along the modified region, at the expense of the estimated European proportion. Presumably, the Native American reference haplotypes in the modified region are now able to approximate reasonably well windows containing both Native American and European haplotypes; in some cases, they will approximate the European windows better than the CEU haplotypes, and hence the increase in the estimated Native American proportion.

Figure 4 shows, for each region and for each fraction of modified (CHB+CHD) haplotypes, the maximal deviation of the average estimated European ancestry $\hat{p}$ from the true average European ancestry p obtained across all loci in the modified region. More precisely, we provide the maximal value of the statistic $D = \frac{|\hat{p}-p|}{\sqrt{p(1-p)}}$, which can be used to obtain the scale of the p-value for testing the null hypothesis of the modified region having the average genome-wide fraction of European ancestry: Given a sample of size N, the p-value is computed as $1-\Phi(D\sqrt{N})$. For example, when the number of modified haplotypes is 30 (0.18 of the Native American panel), the resulting p-value of the most severely affected region

in our simulations (N=200) is $2 \cdot 10^{-2}$, while for a sample of size 1000 we obtain that a similar effect would yield a p-value of $2 \cdot 10^{-6}$. We note that we observe similar but smaller effect when modifying shorter segments; ultimately, for large enough samples and under the assumption of a small finite reference panel, these local biases would appear as statistically significant local deviations in the ancestral proportions. However, for low levels of gene flow ($\leq 6\%$) Figure 4 shows that the biases in local ancestry are unlikely to produce large deviations, and would be statistically significant only at very large sample sizes.

3.6 Assessment of local ancestry performance in real Latinos

In order to estimate the precision of local ancestry inference methods in real data, for which the true local ancestry is unknown, we leverage the fact that local ancestry needs to follow Mendelian inheritance rules. For example, if the father has African local ancestry on both chromosomes while the mother has European ancestry, the child's local ancestry has to have a single chromosome that is African and one that is European. Therefore, pedigree relationships can be used to identify errors in local ancestry estimation by simply testing whether the inferred ancestral status of the child's chromosomes can arise through Mendelian inheritance from the ancestral status of the parent chromosomes. This is done by estimating the local ancestry of each individual in the pedigree separately, and then integrating the trio information to test each genomic position for inconsistency. Any such inconsistency indicates at least one erroneous call in the local ancestry assignments of the trio, so that the counts of the Mendelian inconsistencies in local ancestry (MILANC) give a direct lower bound on local ancestry inference error rate. A critical feature of MILANC is that it is computed without knowing the true ancestry in real data; for this reason LAMP-HAP, which is designed to produce MILANC $= 0$, is not tested in this section.

We first investigated the relation between MILANC and the true underlying error rate. When introducing erroneous calls in the local ancestry of our simulated trios using a random uniform error model, we observed that roughly one third of inserted errors lead to Mendelian inconsistencies, thus indicating that MILANC captures only one component of the true error rate.

Next, we assessed the accuracy of LAMP-LD and WINPOP in empirical data using 232 Mexican and 257 Puerto Rican nuclear mother-father-child families. These trios were collected as part of the Genetics of Asthma in Latino Americans (GALA) Study (Burchard *et al.*, 2004); GALA is a multi-center, international effort designed to identify and directly compare clinical, genetic, and environmental risk factors associated with asthma, asthma severity, and drug responsiveness among Latino ethnic groups. The trios were ascertained on an asthmatic proband. When running local ancestry inference, as proxy for the African (European) ancestry we used the 226 (224) haplotypes of the HapMap 3 phase 2 YRI (CEU) population, while for the Native American ancestry we used 88 Native American samples (25 Bolivian Aymara, 24 Peruvian Quechua, and 39 Mesoamericans) (Bigham *et al.*, 2010). We intersected all SNP sets to achieve a combined panel of 588,595 SNPs.

Table 3 shows the average MILANC attained by WINPOP and LAMP-LD in the GALA trios. We note that the empirical metric of accuracy (MILANC) shows that the accuracy in real data roughly matches the results of our simulations (see Table 1), given that we expect one third of the errors to yield Mendelian inconsistencies. We also note that modeling LD in the form of ancestral haplotypes appears to have a bigger effect for Puerto Ricans rather than Mexicans.

	Mexicans	Puerto Ricans
WINPOP	3.12 (0.03)	3.12 (0.03)
LAMP-LD	3.16 (0.03)	2.50 (0.03)

Table 3. Average genomic MILANC (standard error of the mean) in % attained by best performing methods that model or ignore ancestral population LD in the Mexican and Puerto Rican trios of the GALA study.

4 DISCUSSION

We introduced novel methods for accurate local ancestry inference in multi-way mixtures of populations such as Latinos. Through simulations and analysis of real Latino family data, we demonstrated that our methods attain superior accuracy and scalability compared with current state of the art methods for local ancestry inference. Our methods are implemented as an open source software package for the genetics community. As future work, we mention the incorporation of varying recombination rates into the parameter estimation step of our approach, as well as an adaptive selection of the window length as a function of the genetic distance among ancestral populations at any locus in the genome. In our simulations we have assumed non-population specific recombination rates, however it is straightforward to incorporate population specific recombination rates into our model by using appropriate recombination maps in the training of the ancestral HMMs. Finally, we note that methods for local ancestry inference in multi-way admixed populations are an active area of research (Henn *et al.*, 2012; Johnson *et al.*, 2011; Bercovici *et al.*, 2012). A systematic comparison of the performance of these methods on Latinos as well as admixtures of more than three ancestral populations merits further study.

ACKNOWLEDGMENTS

We thank Lindsey Roth, and Scott Huntsman for helping in the GALA trios QC and genotyping. The authors acknowledge the patients and their families for their participation. The authors also thank the numerous health care providers and community clinics for their support and participation in the GALA Study. We thank Abigail Bigham and Mark Shriver for allowing us to use the Native American data. We would also like to thank Alkes L. Price, Nick Patterson and Noah Zaitlen for helpful comments and suggestions. This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. E.H and Y.B were partially supported by the Israeli Science Foundation, grant no. 04514831, and by the IBM open collaborative research.

REFERENCES

Bercovici,S. *et al.* (2012). Ancestry inference in complex admixtures via variable-length markov chain linkage models. *Proceedings of the 16th Annual International Conference in Computational Biology (RECOMB) (in press)*.

Bigham,A. *et al.* (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet*, **6**(9).

Bryc,K. *et al.* (2010). Genome-wide patterns of population structure and admixture among hispanic/latino populations. *Proc. Natl. Acad. Sci. USA*, **107**(Supplement 2), 8954.

Burchard,E. *et al.* (2004). Lower bronchodilator responsiveness in puerto rican than in mexican subjects with asthma. *Am. J. Respir. Crit. Care Med.*, **169**(3), 386.

Burchard,E. *et al.* (2005). Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am J Public Health*, **95**(12), 2161.

Henn,B. *et al.* (2012). Genomic ancestry of north africans supports back-to-africa migrations. *PLoS Genet*, **8**(1), e1002397.

Hinch,A. *et al.* (2011). The landscape of recombination in african americans. *Nature*, **476**(7359), 170–175.

Hoggart,C. *et al.* (2004). Design and analysis of admixture mapping studies. *Am J Hum Genet*, **74**, 965–978.

Johnson,N. *et al.* (2011). Ancestral components of admixed genomes in a mexican cohort. *PLoS Genet*, **7**(12), e1002410.

Kennedy,J. *et al.* (2008). Genotype error detection using hidden markov models of haplotype diversity. *J Comput Biol*, **15**(9), 1155–1171.

Kimmel,G. and Shamir,R. (2005). gerbil: Genotype resolution and block identification using likelihood. *Proc. Natl. Acad. Sci. USA*, **102**(1), 158–162.

Li,N. and Stephens,M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, **165**(4), 2213–2233.

Mao,X. *et al.* (2007). A genome-wide admixture mapping panel for hispanic/latino populations. *Am J Hum Genet*, **80**(6).

Morales Carrión,A. (1983). Puerto rico: A political and cultural history. *Norton, New York*.

Pasaniuc,B. *et al.* (2009a). Imputation-based local ancestry inference in admixed populations. *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications*, **5542**, 221–233.

Pasaniuc,B. *et al.* (2009b). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, **25**(12), i213–i221.

Pasaniuc,B. *et al.* (2011). Enhanced statistical tests for gwas in admixed populations: assessment using african americans from care and a breast cancer consortium. *PLoS Genet*, **7**(4), e1001371.

Patterson,N. *et al.* (2004). Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, **74**, 979–1000.

Price,A. *et al.* (2007). A genomewide admixture map for latino populations. *Am J Hum Genet*, **80**(6), 1024–1036.

Price,A. *et al.* (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, **5**(6), e1000519.

Reich,D. *et al.* (2005). A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet*, **37**(10), 1113–1118.

Sankararaman,S. *et al.* (2008). Estimating local ancestry in admixed populations. *Am J Hum Genet*, **8**(2), 290–303.

Scheet,P. and Stephens,M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, **78**(4).

Seldin,M. *et al.* (2011). New approaches to disease mapping in admixed populations. *Nat Rev Genet*, **12**(8), 523–528.

Sundquist,A. *et al.* (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res.*, **18**(4), 676–682.

Tang,H. *et al.* (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, **79**, 1–12.

Tang,H. *et al.* (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet*, **81**(3), 626–633.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Tian,C. *et al.* (2007). A genome-wide snp panel for mexican american admixture mapping. *Am J Hum Genet*, **80**(6).

Wegmann,D. *et al.* (2011). Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet*, **43**(9), 847–853.

Yang,J. *et al.* (2011). Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat Genet*, **43**(3), 237–241.

Zhu,X. *et al.* (2004). Linkage analysis of a complex disease through use of admixed populations. *Am J Hum Genet*, **74**(6), 1136–1153.